Tips for Effective Dimensionality Reduction (DR)

Tip 1 → Choose an appropiate method.

- In general, linear method such as PCA are good at preserving global structure.
- Non linear method such as Kernel PCA, t-SNE better at preserving local structure.
- If observations in data have assigned class labels and goal is to represent that best seperates them into known categories, we consider using supervised DR techniques like LDA.

Tip 2 → Preprocess continuous data and count input data.

- Before applying any DR technique, data standardization/scaling is required.
- Scaling step ensures equal contribution from each variable, which is especially important for datasets containing heterogenous features with high variance or distinct units.

Tip 3 → Handle categorical variable appropiately.

- If the dataset contains large set of categorical variable, multiple factor analysis is used.
- PCA cannot be applied to categorical variables, because its objective is to maximize the variance accounted for, a concept that exist only for numerical measure. For categorical variables like "nominal" (unordered) or "ordinal" (ordered) categorical variable, variance can be replaced by chi-square distance on category frequencies.
- Converting categorical variables to dummy binary features, another approach is to use Optimal Scaling PCA. (CATPCA). Optimal scaling replaces original levels of categorical variable with category quantification such that the variance in new variable is maximized. Advantage of optimal scaling is that it does not assume linear relationship between variables.

Tip 4 → Use embedding methods for reducing similarity & dissimilarity input data.

- When features (column name) are not available, relationship between data points measured as similarities / dissimilarities.

- If the original data are binary, Euclidean distance is not appropiate and Manhatten distance is better. If the feature are sparsed, then Jaccard distance is preferred.
- Dissimilarities can be also used as an input to t-SNE.
- A collection of neural network based approaches, called Word2Vec use similarity score (the co-occurence data) to generate vector embedding of objects in a continuous Euclidean space.

Tip 5 → Consciously decide on the number of dimensions to retain.

- Choose number of PCs which explain maximum variance from data.
- For non spectral, optimization based methods, number of components is usually prespecified before DR computation: when using these approach, number of component can be choosen by repeating DR process using an increasing number of dimensions and evaluating whether incoporating more components acheives a significant lower value of loss function that the method minimize eg KL divergence transition probability defined for input & output data in the case of t-SNE.

Tip 6 → Understanding the meaning of the new dimensions.

- Feature map or correlation circles can be used to determine which original variables are associated with each other or with newly generated output dimensions.

Tip 7 → Check the robustness of result & quantify uncertainities.

- When subsequence eigen vectors have close to equal values then PCA is unstable. Then we should check against different parameter setting.
- For t-SNE, use KL divergence to choose optimal perplexity.
- On other hand, if a dataset contains aberrant, robost Kernel PCA should be use